EMBL-EBI Tutorial

Jennifer McDowall (V2.1, Oct 2009)

Getting Started with InterPro

This tutorial provides an introduction to InterPro, the user interface and the database content. Exercises are provided to help you practice what you learn in this tutorial. Further information can be found in the InterPro user manual: http://www.ebi.ac.uk/interpro/user_manual.html.

You will learn about:

- Querying InterPro via the web interface
- Using signatures to classify and annotate protein sequences
- Understanding signature relationship hierarchies
- Relating signatures to protein structure

Contents:

What is InterPro?	2
What are protein signatures?	2
The InterPro home page	2
How do I search InterPro?	2
EXERCISE 1: Searching InterPro	3
EXERCISE 2: General annotation	4
EXERCISE 3: Relationships	5
EXERCISE 4: Structure	6
Further reading	8
Answers to exercises	9

What is InterPro?

InterPro is a curated resource for classifying and annotating protein families, domains and sites. InterPro combines a number of independent protein signature databases (referred to as member databases) into a single searchable resource. InterPro provides protein signatures for over 80% of proteins in the UniProt database. InterPro provides additional information through GO mapping, structural links, external database links and comprehensive abstracts and references.

What are protein signatures?

Protein signatures are obtained by modelling the conservation of amino acids at specific positions within a group of related proteins (family), or within the domains/sites shared by a group of proteins. The different member databases use different computational methods to produce protein signatures:

- Regular expressions (PROSITE patterns)
- Fingerprints that use Position Specific Sequence Matrices (PRINTS)
- Sequence clustering via PSI-BLAST (PRODOM)
- Sequence matrices (PROSITE profiles, HAMAP)
- hidden Markov Models (Pfam, TIGRFAMs, PIRSF, Superfamily, Gene3D, PANTHER, SMART)

These protein signatures are run against the UniProt database of protein sequences, and all significant matches are reported in InterPro.

The InterPro home page

The InterPro home page can be accessed by pointing your web browser at www.ebi.ac.uk/interpro. The home page provides:

- Search tools
- Documentation (user manual, release notes...)
- FTP site for downloads

How do I search InterPro?

InterPro can be searched a number of different ways:

- **Text search**
  - Via the search bar at the top of all InterPro web pages.
  - Search using: UniProtKB accessions; InterPro entry IDs; GO terms; plain text.
- **InterProScan**
  - Link to InterProScan on InterPro home page.
  - Search using protein or nucleotide sequence.
  - Incorporates all the search engines from the member databases.
- **BioMart**
  - Link to BioMart on InterPro home page.
  - Easy querying; retrieve results in HTML, plain text or Microsoft Excel spreadsheet.

Note: To analyse a specific protein, do a text search for its UniProtKB accession number – this will provide both the signature hits and structural information for the protein. By contrast, the InterProScan search only provides the signature hits. InterProScan is best used for novel sequences not found in UniProtKB. If you have an accession number from GenBank, Xref, EMBL or Ensembl, you can convert it to a UniProtKB accession number using PICR (http://www.ebi.ac.uk/Tools/picr/).

Exercise 1:

Part 1: Searching InterPro using BioMart

- Navigate to the InterPro homepage on http://www.ebi.ac.uk/interpro and click on the 'InterPro BioMart' link in the menu on the left.

Let's use BioMart to query an InterPro entry for protein information:

- Create a BioMart search for InterPro entry IPR003121 to answer the following query:

? What proteins match IPR003121? Provide a list the UniProtKB protein accessions and UniProt Ids (name), the source protein database, their match scores, and their match start and stop positions?

BioMart is also linked to other databases, namely Reactome and PRIDE. Keeping the existing query, let's use it to retrieve additional information from Reactome:

- Amend the previous query by linking it to the Reactome reaction database.
- Limit the Reactome reactions to only those in *Homo* sapiens, and restrict it to unique results only. Use BioMart to answer the following query:

? What Reactome reaction stable ID and Reaction name correspond to the unique human proteins matched by IPR003121?

We have performed these searches querying a single protein or a single InterPro entry. However, we could also submit a list of proteins or entries, either by typing in a comma-separated or space-separated list, or by uploading a file containing such a list.

Part 2: Searching InterPro using a text search

- Open the InterPro homepage (http://www.ebi.ac.uk/interpro/) in a web browser.
- Using the InterPro Search box on the top right of the page, type in the UniProtKB accession '**O15075**'.

You should now have a picture of the signature matches for the above protein.

- Read below to understand how to interpret this protein match:

How do I interpret an InterPro protein graphical view?

In InterPro, the graphical view of a protein consists of the following features:

- **InterPro signature**
  - Each solid coloured bar represents a signature that matches the protein.
  - The colour of the bar denotes the member database (e.g. dark blue for Pfam).
  - If you hover your mouse over the coloured bar, a pop-up will display the signature accession and the sequence matched (note: this only works in some browsers).
  - The length of the protein is indicated by the white vertical bars, which are marked every 10 amino acids.
  - The InterPro entry the signature belongs to is marked on the left-hand column (note: several signatures can belong to the same InterPro entry (look at IPR003533) if they are redundant, representing the same sequence in the same set of proteins).
  - Unintegrated signatures are listed at the bottom (these have not yet been curated).
- **Structural Features**
  - The green striped bar represents coverage of PDB chain (PDB accession on the left).
  - The pink- and black-striped bars represent the structural classification databases CATH and SCOP, respectively.
  - CATH and SCOP break the PDB into its structural domains.
- **Structural Predictions**
  - The yellow- and red-striped bars represents the homology databases ModBase and Swiss-Model, respectively.
  - These homology models estimate the structure based on the closest homologue.

? Looking at the InterPro graphical results for O15075, how many InterPro entries (not individual signatures) match the query protein sequence?

? How many domains is the protein be divided up into?

Hint: There can be more than one InterPro entry for a single domain, because InterPro represents domain hierarchies (i.e. structural domain superfamilies, domain families and domain subfamilies). Therefore, look at all the signatures together, and see how many regions the protein is broken into along the sequence.

Exercise 2:

Exploring InterPro Entries: General Annotation

Let's start with the top entry, IPR000719. Notice that there is a second entry, IPR011009 that covers the same sequence position on the protein. We will see late that this represents the same sequence in a broader set of proteins.

☞ Click on the hyperlink to IPR000719 in the left-hand column.

? What is the name of this domain?

Scroll down to the "Signatures" section.

This section lists the signatures in an entry, the database they come from, and the number of proteins they match.

? Which signatures make up this entry and how many proteins are matched?

Scroll down to the "GO term annotation" section.

InterPro provides its own mappings to GO terms based on the curated UniProt/Swiss-Prot proteins matching an entry. These are useful for TrEMBL proteins that do not otherwise have GO terms associated with them.

? What GO terms does this entry provide?

Scroll down to the "Taxonomic coverage" section (taxonomic wheel).

InterPro divides all the protein hits in an entry by their taxonomy.

? How wide a taxonomic coverage do proteins containing a protein kinase domain have?

Exercise 3:

Exploring InterPro Entries: Relationships

Scroll to the "InterPro Relationships" section of the IPR00719 entry.

InterPro links related signatures through relationships: Parent/Child relationships indicate domain/family hierarchies, and Contains/Found in relationships subdivides domains/families into sequence regions.

? What "Child" entries are IPR000719 subdivided into?

Child entries subdivide IPR000719 into more closely related subgroups.

? What is the name of the "Parent" of IPR000719?

The parent entry represents domains with a structural fold homologous to that of the protein kinase domain (even if they have no enzyme activity), whereas IPR000719 represents proteins more closely related by sequence.

There are too many entries that IPR000719 is Found In to list them, but take a look at them. Most of these entries are families of related proteins that all contain a protein kinase domain.

Note: only a few of these relationships mentioned above actually apply to our protein.

Exercise 4:

Exploring InterPro Entries: Structure

- Return to the InterPro graphical view page of our protein, O15075.
- Scroll down to the "Structural features" just below the view of signatures.

The green-striped bar represents the PDB structure, and its length indicates the region of the protein for which the structure is known. The pink-striped bar represents the CATH database and the black-striped bar the SCOP database, both of which are structural classification databases that break PDB structures down into their constituent domains for classification.

? What region is covered by the PDB structure (ie which domain)?

Hint: What is the name of the entry IPR003533, and does it cover both domain hits that the signatures in IPR003533 cover?

Not all of the protein has been structurally characterised, shown by the fact that only a small region of this protein is covered by the PDB green-striped bar. To help address this problem, there are homology models from both ModBase (yellow striped bar) and Swiss-Model (red-striped bar) found under the "Structural Predictions" section. These are models based on aligning our protein with its closest homologue whose structure has been determined. *Note: these are predictive models that only provide a 'best guess' at the remaining structure.*

Let's have a look at the structure of the doublecortin domain in O15075 using AstexViewer®. You will note that there is an AstexViewer® symbol ( ) beside each CATH and SCOP-classified domain structure. Each structural view will give you the entire PDB displayed in green with the particular domain chosen highlighted in yellow. However, with O15075, there is only one domain to view, therefore the entire domain should be coloured yellow. There are 2 accession numbers next to each symbol:

(1) 3.10.20.230 and d.15.11.1 refer to the CATH and SCOP accessions, respectively.
(2) 1mg4A00 and d1mg4a refer to the PDB the picture is derived from (PDB 1mg4, chain A).

- Click on the symbol adjacent to the N-terminal SCOP domain (d.15.11.1). (The image will pop-up in a new window).

You should now have a "Cartoon" view of the doublecortin domain of the O15075 protein (PDB 1mg4, chain A).

The image can easily be rotated:

- Place the mouse over the image, and by holding down the left mouse button down, use the mouse to rotate the structure to any angle you want in order to get a better view.

? What is the predominant topology of this protein, alpha helix, beta sheet, or both?

Note: beta-sheets are displayed as an arrow, while alpha helices are displayed as a coil.

There are some ligands associated with this structure. To more easily view them, change their view to a solid surface:

- Click on "Ligand" in the left-hand column. From the drop-down menu, click on "Solid Surface".

? How many ligands are displayed?

The residues that bind each ligand can be explored further. It is easiest to view the protein in the "Line" view:

- Click on "Protein" in the left-hand column. From the drop-down menu, click on "Line".

Your protein will now have both the "Cartoon" view and the "Line" view. To remove the "Cartoon" view:

- Click on "Protein", and click on "Cartoon" in the drop-down menu to de-select it.

We can now highlight the residues that contact the ligand:

- Click on "Chemistry" in the left-hand column. Click on "1,1mg4:SO4".

The pop-up box describes the residues that bind to this ligand:

- Click on one of the residues in the pop-up chemistry box.

Your AstexViewer should zoom in to the correct residue and highlight it in bold:

- Click on "Zoom Out" button in the left-hand column.

The highlighted residue should remain in bold.

Now let's try the structure-sequence interactive display. Keep the protein in the "Line" view, as it allows specific residues to be selected more easily.

- Look at the sequence for this structure at the foot of the Astex window.

To move along the sequence, simply hold the cursor in the lower section that contains the sequence. Move the mouse left to scroll towards the C-terminus and right for the N-terminus.

- Hover the mouse over the first amino acid in the sequence

The 3-letter amino acid code for the residue, as well as its position, will appear as a footnote.

- Click on any residue.

The image will zoom in to that residue in the 3-D image.

- Use the "zoom out" button to return the structure to its full view.

Note: the residue is still highlighted once you zoom out.

- Click on any residue on the structure. Click on "zoom out".

Note: the residue in the sequence is highlighted, and it is marked in the structure.

- Close the Astex window. Go back to the Graphical View of O15075.

This is the end of the short tour of the InterPro database, available at the EBI. Perhaps you might like to try it again with a more relevant sequence.

Further reading

InterPro: the integrative protein signature database. Hunter S, *et al.* Nucleic Acid Res (2009) **37**:D211-5.

InterPro and InterProScan: tools for protein sequence classification and comparison. Mulder NJ, Apweiler R. Methods in Molecular Biology (2007) **396**:59-70.

New developments in the InterPro database. Mulder NJ, *et al.* Nucleic Acids Research (2007) **35**:D224-8.

Answers to Exercises

Below are the answers to the tutorial questions, but you will learn more if you try the questions yourself first.

Exercise 1:

Part 1: Searching InterPro using BioMart

? What proteins match IPR003121? Provide a list the UniProtKB protein accessions and IDs, the source protein database, their match scores, and their match start and stop positions?

UniProtKB Protein Accession	UniProtKB Protein ID (Name)	Source Protein Database	Match Score	Match Start Position	Match Stop Position
A0CSW8	A0CSW8_PARTE	UniProt/TrEMBL	8.20001e-19	182	267
A0EDG0	A0EDG0_PARTE	UniProt/TrEMBL	1.7e-18	192	271
A0NDQ1	A0NDQ1_ANOGA	UniProt/TrEMBL	4.6e-25	191	272
A1A5J8	A1A5J8_XENLA	UniProt/TrEMBL	4.6e-30	283	362
A1CR41	A1CR41_ASPCL	UniProt/TrEMBL	3.49999e-19	273	359
A1CS93	A1CS93_ASPCL	UniProt/TrEMBL	5.99999e-24	200	287
A1D3Z6	A1D3Z6_NEOFI	UniProt/TrEMBL	7.6e-20	275	361
A1D559	A1D559_NEOFI	UniProt/TrEMBL	2.1e-24	200	287
A1TR45	A1TR45_ACIAC	UniProt/TrEMBL	4.1e-24	19	100
A1TW24	A1TW24_ACIAC	UniProt/TrEMBL	1.6e-21	896	982

Above are the top 10 hits only.

➢ Search settings:

- Dataset = *InterPro BioMart*; *InterPro Entries*
- Filters = under "InterPro Entry ID" in the 'InterPro Entry Filters' section enter "IPR003121" (Note: make sure it is also checked)
- Attributes = check *UniProtKB Protein Accession*, *UniProtKB Protein ID*, *Source Protein Database*, *Match score*, *Match Start Position* and *Match Stop Position* (under "Protein Matches")

? What Reactome reaction stable ID and Reaction name correspond to the unique human proteins matched by IPR003121?

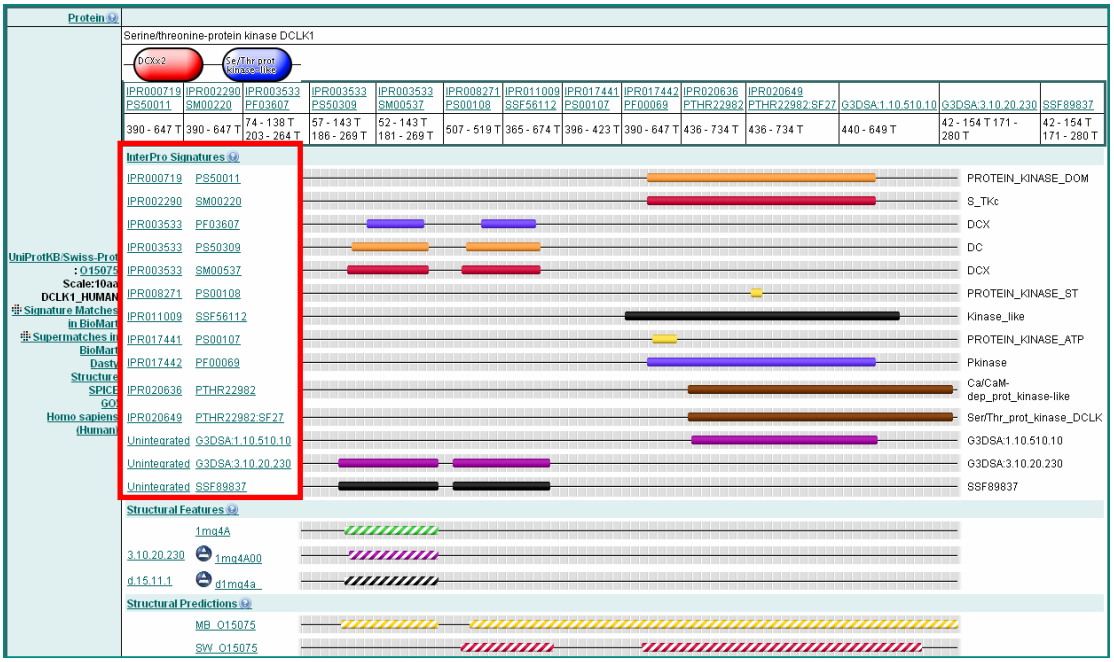
UniProtKB Protein Accession	UniProtKB Protein ID (Name)	Source Protein Database	Match Score	Match Start Position	Match Stop Position	Reaction stable ID	Reaction name
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	0	17	123	REACT_988	Phosphorylation of MDM2 at serine-395 by ATM kinase
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	1.9e-22	27	102	REACT_988	Phosphorylation of MDM2 at serine-395 by ATM kinase
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	8.1e-28	189	284	REACT_988	Phosphorylation of MDM2 at serine-395 by ATM kinase
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	1.29998e-41	25	112	REACT_988	Phosphorylation of MDM2 at serine-395 by ATM kinase
Q6STE5	SMRD3_HUMAN	UniProt/Swiss-Prot	1.9e-26	259	338	REACT_19175	Activation of PPARA by Fatty Acid Ligands
Q6STE5	SMRD3_HUMAN	UniProt/Swiss-Prot	1.2e-25	259	334	REACT_19175	Activation of PPARA by Fatty Acid Ligands
Q6STE5	SMRD3_HUMAN	UniProt/Swiss-Prot	4.3e-28	252	344	REACT_19175	Activation of PPARA by Fatty Acid Ligands
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	0	17	123	REACT_12537	AKT phosphorylates MDM2
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	0	17	123	REACT_18393	Trafficking of GluR1-containing AMPA receptors
Q00987	MDM2_HUMAN	UniProt/Swiss-Prot	1.9e-22	27	102	REACT_12537	AKT phosphorylates MDM2

Above are the top 10 hits only.

➢ Search settings as above, plus:
- Dataset = REACTOME (CSHL) reaction
- Filters = select *Homo sapiens* under "Limit to Species"
- Attributes = check *Reaction stable ID* and *Reaction name* (under "Reaction")
- On results page, check "Unique results only"

Part 2: Searching InterPro using a text search

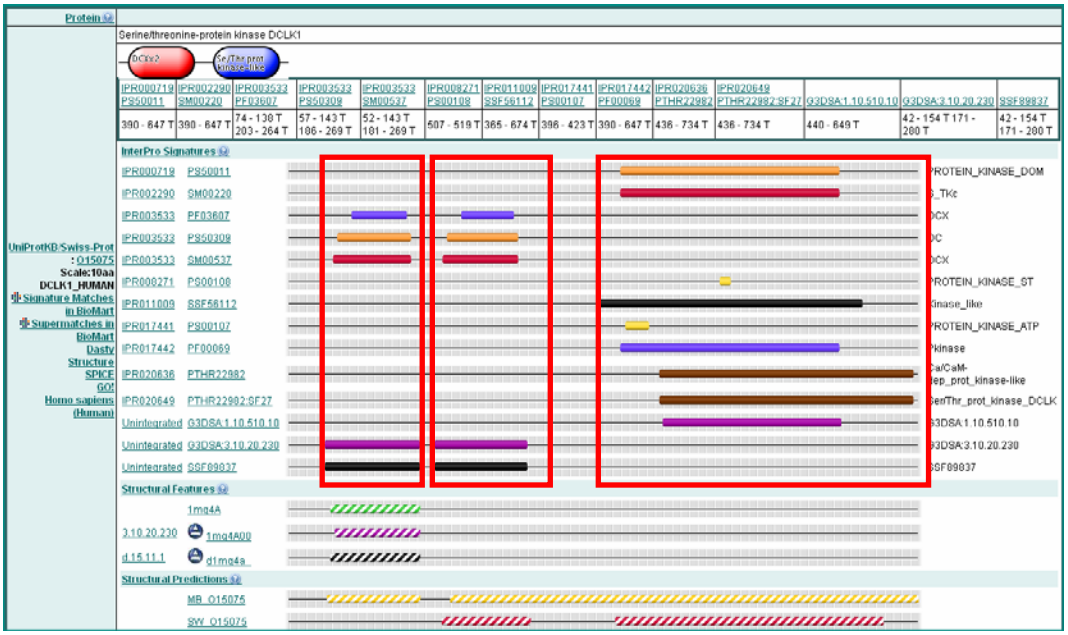
? Looking at the InterPro graphical results from O15075, how many InterPro entries (not individual signatures) match the query protein sequence?



➢ From the list of InterPro accession numbers, you can see that there are 9 different entries that match O15075 (IPR000719, IPR002290, IPR003533, IPR008271, IPR011009, IPR017441, IPR017442, IPR020636, IPR020649). Some of these entries contain more than one signature, for example, IPR003533 contains PF03607, PS50309 and SM00537.

(Note that the bottom three signatures are marked as "unintegrated". These signatures have not yet been curated, but their results are still presented).

? How many domains is the protein be divided up into?



- From the picture above, there are:
  - 2 matches to IPR003533 (Doublecortin domain)
  - 1 match to IPR000719 (Protein kinase catalytic domain) (covers same region as IPR002290/IPR011009/IPR017442/IPR020636/IPR020649)

Exercise 2:

Exploring InterPro Entries: General Annotation

? What is the name of this entry?

- Protein kinase catalytic domain.

? Which signatures make up this entry and how many proteins are matched?

- Only one: PS50011, matching 64392 proteins.

? What GO terms does this entry provide?

- Three GO terms are provided:
  - Process = GO:0006468 (protein amino acid phosphorylation)
  - Function = GO:0004672 (protein kinase activity)
  - Function = GO:0005524 (ATP binding)

? How wide a taxonomic coverage do proteins containing a protein kinase domain have?

- Ubiquitous.

Exercise 3:

Exploring InterPro Entries: Relationships

? What "Child" entries is IPR000719 subdivided into?

➢ There are 3 Child entries:
- IPR001245 (Tyrosine-protein kinase, catalytic domain)
- IPR017442 (Serine/threonine-protein kinase-like domain)
- IPR018934 (RIO-like kinase)

? What is the name of the "Parent" of IPR000719?

➢ Protein kinase-like domain (IPR011009)

Exercise 4:

Exploring InterPro Entries: Structure

? What region is covered by a PDB structure (ie which domain)?

➢ The PDB only covers the first doublecortin domain of O15075.

? What is the predominant topology of this protein, alpha helix, beta sheet, or both?

➢ SCOP describes the topology of this domain as *beta(2)-alpha-beta(2)*, therefore it is a mixture of beta-sheets and an alpha-helix.

? How many ligands are displayed?

➢ Two.